COMMENTS ON MULTIPLE COMPARISONS PROCEDURES

by

W. T. Federer

Panel Discussion on "The Use of Mean Separation Techniques on n x n Set of Means, Interaction Means Resulting From a Cross-Classified Analysis of Variance", Joint Statistical Meetings, August 27, 1975, 8:00 p.m., Atlanta, Georgia, Chairman: Judson U. McGuire, Jr.

BU-577-M in the Mimeo Series of the Biometrics Unit, Cornell University Ithaca, N. Y. 14853.

COMMENTS ON MULTIPLE COMPARISONS PROCEDURES

by

W. T. Federer

BU-577-M January, 1976

ABSTRACT

Similarities of fixed range multiple comparisons procedures are discussed along with four types of error rate bases. Mention is made of some variable range procedures, some multiple F-test procedures, and some combined F-test and multiple range procedures. Then attention is focused on multiple comparisons in a split plot design for the entire set of comparisons and for the subset eliminating comparisons among whole plot means. A fixed range analogue of Duncan's new multiple-range procedure (variable range) is described. Then it is shown how to construct an infinite class of multiple range procedures (with unknown properties). Some comments on preliminary F-tests prior to use of a multiple comparisons procedure are presented. Some suggested reading for additional information is given.

COMMENTS ON MULTIPLE COMPARISONS PROCEDURES

by

W. T. Federer

BU-577-M January, 1976

In starting out, I would like to make certain that everybody understands the items we are discussing. I think quite often that a lot of problems people have are much ado about nothing other than differences in definition. Also, many times one can do a quick approximation and have essentially all the nice properties needed.

First, we should understand the error rate basis and secondly, we should understand the computational procedures for the different multiple range procedures, and observe their similarities. A description of many of the multiple range procedures are described in the citations given at the end of this paper under "Suggested Reading". Four error rates considered here are (see Federer [1964] for definitions):

α_s = error rate per comparison vs. comparisonwise error rate.

α_e = error rate per experiment.

α_h = experimentwise error rate.

α_p = degree-of-freedomwise error rate.

There are more bases for setting error rates than these (e.g. see Tukey [1953] and Miller [1966]). Now for experiments with equal replication on the treatments we first compute a standard error of the mean as

$s_{\bar{y}} = \sqrt{\text{error mean square}/\text{no. of replicates} = r}$ with f degrees of freedom. Simultaneous confidence intervals on differences between a pair of means are computed as follows:

lsd procedure (error rate base is comparisonwise or per comparison)

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm t_{\alpha_s, f}\sqrt{2}s_{\bar{y}}$$

where $t_{\alpha_s, f}$ is the tabled value of the α_s percent level t-statistic for f degree of freedom, $\bar{y}_{i.}$ is a treatment mean obtained from the experiment, $i = 1, 2, \cdots, t$ treatment, and $1 - \alpha_s$ is the confidence coefficient. The lsd procedure is sometimes called a least significant difference. Way back in the '30's and '40's α_s was 5% for the lsd and α_s = 1% was called the most significant difference which we could label as msd. As we use the lsd today, α_s could be any value between 0 and 1.

esd procedure (error rate per experiment)

Let $\alpha_e = \alpha_{s/m}$ where m is the number of comparisons to be made in an experiment. Then the $1 - \alpha_e$ confidence interval on the true mean difference using a per experiment error rate is computed as:

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm t_{\alpha_e, f}\sqrt{2}s_{\bar{y}} \quad .$$

hsd (or honestly significant difference) procedure (experimentwise error rate for all pairs of means)

For t treatments, $t(t-1)/2$ comparisons among the means, and an experimentwise error rate based of α_h, the $1 - \alpha_h$ confidence interval on the differences between the true means of the treatments is computed as:

$$\bar{y}_{i} - \bar{y}_{i'} \pm q_{\alpha_h, t, f}\, s_{\bar{y}}$$

where $q_{\alpha_h, t, f}$ is the tabled value of the studentized range for v treatments, f degrees of freedom in $s_{\bar{y}}$, and for the α_h percentage point.

Scheffé's procedure or ssd (experimentwise error rate for all possible contrasts among means)

The $1-\alpha_q$ confidence interval on any contrast among t means is computed as:

$$\sum_{i=1}^{t} c_i \bar{y}_i \pm S\, s_{\bar{y}} \sqrt{\Sigma c_i^2}$$

where $S^2 = (t-1)\, F_{\alpha_q}\,(t-1, f) \sum_{i=1}^{t} c_i^2$ and F_{α_q} is the tabulated value of F for $t-1$ degrees of freedom in the numerator and f in the denominator at the α_q percent level.

Dunnett's procedure for comparison with a control (experimentwise error rate)

If one of the t treatments is a control and one wants only comparisons with the control and an experimentwise error rate base, then the $1-\alpha_c$ confidence intervals on comparisons of the other $t-1$ means with the control would be computed as:

$$\bar{y}_c - \bar{y}_i \pm t^*_{\alpha_c, f} \sqrt{2} s_{\bar{y}}$$

where $t^*_{\alpha_c, f}$ is the tabulated value in Table 2.1 of Federer [1964] for f degrees of freedom in $s_{\bar{y}}$ and $i = 1, 2, \cdots, t-1$.

Other fixed range procedures would follow a similar computational procedure which is essentially some constant from the desired table times $s_{\bar{y}}$. A significance test would be obtained by noting whether or not zero fell in the computed confidence interval.

For unequal numbers there are various approximations one can use without disturbing the properties of the procedure too much. (See writings of D. B. Duncan and Federer [1964], pages 8-10.)

One of the main items to understand in the above is that the α's differ, that is the error rate base varies from procedure to procedure. This is why a subscript was put on the α. One should select the error rate base desired and use the procedure achieving this goal. Also note that if an experimenter desires only m comparisons, where $m \neq t(t-1)/2$ for t means, and if one wants an experimentwise error rate, this can be achieved approximately by using the esd in place of the hsd, since the esd and hsd are essentially equal for $m = t(t-1)/2$. Thus, one would be in the right ball park for whatever m one selected. Another approximate procedure would be to use the hsd procedure but to pick t^* such that $t^*(t^*-1)/2$ was approximately equal to m and then to use $q_{\alpha_h, t^*, f}$ in computing the "hsd". Either of the approximations should work fairly well. If you don't like approximations, I should point out that in the Real World you don't have normality, you more than likely don't have equal variances, perhaps you don't have independence, and perhaps you don't have a linear response model. But, you assume that you do and hence are approximating the true situation most of the time in the Real World. In the Classroom World, or the Chalkboard World of Statistics classes, these assumptions can't help but hold but in the Real World they are only approximations. Another approximation is being suggested when the unequal replication situation prevails and when one wants an experimentwise error rate.

Interest should center on what Tukey defined as error rate bases. Decide the base and kind of error rate one desires and then select the approximate procedure. If some editor decides one must use procedure X to publish in his journal, I believe that this is the worst possible editorial policy. Some years ago, an agency, I believe it was TVA, was going to contract with Cornell University to do some research but they indicated that the researcher must use Duncan's New Multiple Range Test procedure in his analysis of the data. The

Department involved came to me and asked why this was even being considered, and I told them to tell TVA to "go jump in the creek" since they were capable researchers and should perform their research and analysis in the best and most appropriate way known to them without having to use some prescribed statistical analysis which may or may not be appropriate. They did this and got the money without the stipulation. For the research they did it was highly unlikely that they would have wanted to use the stipulated procedure.

Sometimes a statistician believes that one should pick a procedure giving fewer "statistically significant" results and then push for a procedure with a larger confidence interval, e.g. the hsd confidence interval is larger than the lsd. The reason given is that individuals, say in psychology, only publish "significant results". Thus if the null hypothesis were really true, selection of a larger interval would result in fewer Type I errors. But now really, why not pick the procedure with the largest confidence interval (Scheffé's) or why not use 6 hsd as the interval? If the user realizes what is going on and what is wanted in terms of the error rate base, the large interval argument won't hold water.

We have been talking about fixed level multiple range procedures. A number of the procedures involve variable ranges which depend upon the number of ranked means between the two means being compared. Some of these are:

- Duncan's New Multiple Range
- Student-Newman-Keuls
- An extension of Scheffé's ssd

These procedures are illustrated in Federer [1964]. There are also multiple F-test procedures and combined F-test and multiple range procedures. Some of these are:

- Hartley and Ghosh multiple F-tests
- Duncan's Multiple Comparisons
- Tukey Gap, Straggler, Variance
- F-test followed by a Multiple Range (either fixed or variable) Procedure. (This is sometimes called a "protected" Multiple Range Procedure.)

Now after these preliminaries, let's get to designs like the split plot which was one of the items J. U. McGuire, Jr., wanted us to consider. Suppose one has a split plot design. In order to simplify my discussion let us not worry about the experiment design for the whole plot treatment, except that we want an orthogonal one for the discussion, and about comparisons among the whole plot treatment means. Let us only consider a split plot design wherein the b split plot treatments have been randomly allotted to the b split plot experimental units. Then we would have a analyses of variances for the a whole plot treatments as follows:

Source of variation	Degrees of freedom for levels of whole plots				Sum
	a_1	a_2	a_3	a_a	
Total	rb	rb	rb	rb	arb
Correction for mean	1	1	1	1	a
Blocks = B	r-1	r-1	r-1	r-1	a(r-1)
Treatments = T (Split plot)	b-1	b-1	b-1	b-1	a(b-1)
B X T	(r-1)(b-1)	(r-1)(b-1)	(r-1)(b-1)	(r-1)(b-1)	a(r-1)(b-1)

If the B X T mean squares all estimate the same parameter and represent estimates of the error variance, then one could use a pooled variance with a(r-1)(b-1) degrees of freedom; this pooled mean square is the error (b) in the standard textbook analysis of variance for split plot designs (see Federer [1975]). Looking

at the problem in this manner, one could use an experimentwise error rate for all ab treatments, or for only the b treatments within each whole plot treatment. For the former case use the following procedure:

$$\bar{y}_{i.} - \bar{y}_{i.'} \pm q_{\alpha_h, ab, fa} \, s_{\bar{y}w} , \tag{1}$$

where $\bar{y}_{i.}$ are whole plot treatment means and $s^2_{\bar{y}w}$ equals error (a) mean square $(E_a)/rb$;

$$\bar{y}_{ij} - \bar{y}_{ij'} \pm q_{\alpha_h, ab, fb} \sqrt{\text{error (b) mean square } (E_b)/r} \tag{2}$$

where $fb = a(r-1)(b-1)$ degrees of freedom and $j \neq j'$; and

$$\bar{y}_{ij} - \bar{y}_{i'j'} \pm q_{\alpha_h, ab, f^*} \, s^*_{\bar{y}} \tag{3}$$

where $i \neq i'$ (i.e. different whole plots), j may or may not be equal to j', $s^*_{\bar{y}} = \sqrt{\left(\frac{E_a + (b-1)E_b}{rb}\right)}$, and f^* is the approximated value for degrees of freedom in $s^*_{\bar{y}}$.

If one only desires to compare means of split plot treatments within whole plot treatments, and to have an experimentwise error, then only the last two formulas above are appropriate with $v^* = a(b-1)$ replacing ab in the $q_{\alpha_h, ab, fb(\text{or } f^*)}$ term.

On a different issue, consider an approximation which I use in place of Duncan's new Multiple Range Procedure (a variable range procedure). It involves

(i) ranking the means

(ii) computing $s_{\bar{y}}$ where $d_{\alpha_p, p, f}$ is the tabled value of the statistic at the α_p percent level (e.g. see Table II.3 in Federer's Experimental Design or Harter's tables) and where $2 \leq p \leq t$.

(iii) comparing the range of p means with the value in (ii) I use a fixed range procedure which involves only the term for p = t, i.e. $d_{\alpha_p, t, f}\, s_{\bar{y}}$. Dave Duncan may not like this but I don't think that the properties of the test are changed to any appreciable degree. It certainly is a simple first approximation. I infer this is okay from Hartley's 1955 paper where he uses what he calls a sequential F-test. Suppose there are m contrasts in an analysis of variance on which one wants to perform F-tests. Hartley first ranks the mean squares according to their respective tail areas in the F-distribution and then obtains values $F_{\alpha/m}$, $F_{\alpha/(m-1)}$, $F_{\alpha/(m-2)}$, $\cdots$, F_{α} where the various F-values correspond to the degrees of freedom in the numerator and the denominator associated with the F-statistic. As soon as non-significance is reached the testing stops. Now Ghosh [1955] showed that if one tested m variances against $F_{\alpha/m}$ that the error rate was α percent experimentwise. Hartley's [1955] sequential F is almost experimentwise. Thus, from this we project the fact that using only $d_{\alpha_p, t, f}\, s_{\bar{y}}$ instead of the variable ranges would be almost a degree-of-freedomwise error rate procedure and, of course, much simpler to use.

Thus, all the extra work involved in multiple variable range procedures can be eliminated without essentially changing the properties of the procedure. The small changes don't matter that much. Furthermore, the added difficulties in teaching variable range procedures is not worth the extra effort. I would rather teach only fixed range procedures such as the lsd, hsd, esd, Scheffé's, Dunnett's, and one proposed by Kurtz, Link, Tukey, and Wallace [1965a,b] which is called a shortcut to allowances procedure. It is called the rsd in Federer [1955, 1964] but it might as easily be labeled scap.

There are several additional procedures possible. One could just stand here and give them out as wanted. For example, one could construct the value $1 + \epsilon$, $0 \leq \epsilon \leq \infty$, and multiply the $t_{\alpha,f}$ statistic by $1 + \epsilon$ to obtain $t_{\alpha,f}(1+\epsilon)\sqrt{2}\, s_{\bar{y}}$ and thus construct an infinite class of procedures. Of course, I don't know the properties of each procedure for each $\epsilon > 0$ but this just illustrates how one can cook up additional multiple range procedures at will. For example, here is another significance multiple range testing procedure:

(i) compute all possible differences between the t means.

(ii) rank all the $t(t-1)/2$ differences.

(iii) compute the lsd.

(iv) compute $\alpha t(t-1)/2 = k$, say.

(v) compare all differences with the lsd, and then all but the k smallest differences larger than the lsd are declared significant.

This is another one of a multitude of possible multiple range test procedures but again we don't know its properties.

In summary, I personally use the lsd or the hsd, depending upon what error rate base I wish to work from. Sometimes I use the esd for situations involving m comparisons.

There are statisticians around who recommend using an F-test before one uses a multiple range testing procedure. They call this a protected multiple range procedure. I would like to hear D. B. Duncan's comments on this. If I read him correctly, the initial F-test is of little or no use if the number of treatments t is moderately large, say greater than 10-12. If one is constructing simultaneous confidence intervals, then it would be pointless to use an initial F-test anyway. It could only be used when one was significance testing.

There is one situation wherein the F-test can become relatively very large but no difference between pairs of means will be significant. Suppose that t/2 of my means fall at one point, say a, and the other t/2 of my means fall at a second point b; then the largest difference between any pair of means is b - a. This difference may be an lsd/10, for example. However, if we let t get large enough, the F-test will quickly exceed significance points such as .05 or .01. The only significant contrast is the t/2 means at a versus the t/2 means at point b. This single degree-of-freedom contrast for t large is the cause of the significant F. Examples for the reverse situation can easily be constructed.

These examples were presented to illustrate that one can get into "trouble" or "conflicts" with any procedure if one doesn't know the characteristics of the procedure. One should really decide what his goals and error rate base are and then proceed accordingly. The use of the so-called protected F-test and multiple range procedure offers little or no protection, whatever that is.

If one wants to learn about the procedures, some literature citations are listed in the following pages. It is too bad that the dittoed material prepared by Tukey in 1953 was not made generally available. However, there is other material available. The paper by Hartley [1955] is worthwhile reading. Duncan has several papers out. Some of his VPI technical reports were very good on particular aspects. I found Harter's [1957] paper very enlightening. The attached mimeo should be very helpful. It was presented in a Summer Session at Colorado State University in 1964 and has been presented since that time to classes at Cornell University. As presented one only requires one or two lectures to teach multiple comparisons procedures. Also, the paper by Kurtz et al. [1965a] is enlightening reading.

Duncan, D. B. [1955]. Multiple range and multiple F-tests. Biometrics 11:1-42.

Federer, W. T. [1955]. Experimental Design--Theory and Application, Macmillan Company, New York, N. Y., pp. xix + 544 + 47. Second printing 1963 (out of print). Indian edition 1967, Oxford & IBH Publishing Company, Calcutta, India (available outside India, Ceylon, Pakistan and Burma only through author).

Federer, W. T. [1964]. Chapter II - Error rates in experiments and multiple range tests. Mimeographed material from Colorado State University and NSF 1964 Summer Session (attached).

Federer, W. T. [1975]. The misunderstood split plot. Applied Statistics, Proceedings of Conference at Dalhousie University, Halifax, May 2-4, 1974, R. P. Gupta, editor, North Holland Publishing Co., Amsterdam, pp. 9-39.

Federer, W. T. and L. N. Balaam [1972]. Bibliography on Experiment and Treatment Design Pre-1968, published for the International Institute by Oliver and Boyd, Edinburgh, Scotland, 769 pp. (see references under Category A3 for pre-1968 literature on multiple comparisons procedures).

Ghosh, M. N. [1955]. Simultaneous tests of linear hypotheses. Biometrika 42: 441-449.

Harter, H. L. [1957]. Error rates and sample sizes for range tests in multiple comparisons. Biometrics 13:511-536.

Hartley, H. O. [1955]. Some recent developments in analysis of variance. Communications of Pure and Applied Mathematics 8:47-72.

Kurtz, T. E., R. F. Link, J. W. Tukey, and D. L. Wallace [1965a]. Short-cut multiple comparisons for balanced single and double classifications: Part 1, Results. Technometrics 7(2):95-161 with discussion and related material on pages 163-262.

Kurtz, T. E., R. F. Link, J. W. Tukey, and D. L. Wallace [1965b]. Short-cut multiple comparisons for balanced single and double classifications: Part 2, Derivation and approximations. Biometrika 52(3 and 4):485-498.

Miller, R. G. [1966]. Simultaneous Statistical Inference, McGraw-Hill Book Co., New York, N.Y., pp. xv + 272 (with a bibliography).

Tukey, J. W. [1953]. The Problem of Multiple Comparisons, Unpublished dittoed notes, Princeton University, Princeton, N.J., 396 pages.